Predikcia srdcových ochorení pomocou logistickej regresie

Úvod

Tento projekt sa zameriava na **predikciu srdcových ochorení** pomocou **logistickej regresie**. Dataset pochádza z **UCI Machine Learning Repository**, konkrétne verzia dostupná na Kaggle (https://www.kaggle.com/datasets/ketangangal/heart-disease-dataset-uci). Cieľom je analyzovať dostupné zdravotné dáta pacientov a vytvoriť model, ktorý dokáže predpovedať, či daná osoba trpí srdcovým ochorením.

Dataset

Dataset obsahuje viaceré premenné, ktoré opisujú zdravotný stav pacienta. Kľúčové premenné zahŕňajú:

- **age** – vek pacienta
- **sex** – pohlavie pacienta
- **trestbps** – pokojový krvný tlak
- **chol** – hladina cholesterolu v krvi
- **fbs** – hladina cukru v krvi nalačno (>120 mg/dl)
- **restecg** – výsledky elektrokardiografického vyšetrenia
- **thalch** – maximálna dosiahnutá srdcová frekvencia
- **oldpeak** – depresia ST segmentu pri zaťažení
- **slope** – sklon ST segmentu pri zaťažení
- **num** – cieľová premenná (0 = zdravý, 1-4 = rôzne stupne ochorenia)

Predspracovanie dát

Pred samotným tréningom modelu bolo potrebné vykonať viacero krokov:

1. Kontrola a čistenie dát

- Skontrolovali sme **chýbajúce hodnoty** a nahradili ich mediánom (pre číselné premenné) a modusom (pre kateggorické premenné).
- Skontrolovali sme **duplicity.** V datasete sa nenachádazjú.

2. Transformácia premenných

- Kategorické premenné (napr. **pohlavie, typ bolesti na hrudi, ST segment**) boli prevedené na **one-hot encoding**.
- Cieľová premenná **num** bola pôvodne **multikategórická (0-4)**, preto sme ju previedli na **binárnu klasifikáciu (0 = zdravý, 1 = chorý)**.

3. Škálovanie dát

- Všetky číselné premenné boli **škálované pomocou StandardScaler**, aby sa zabezpečila rovnaká váha rôznych atribútov.

Tréning modelu

Použili sme **logistickú regresiu** ako základný klasifikačný model. Dáta boli rozdelené na **80 % tréningová sada, 20 % testovacia sada**.

Použitý kód:

```
from sklearn.model_selection import train_test_split

from sklearn.preprocessing import StandardScaler

from sklearn.linear_model import LogisticRegression

from sklearn.metrics import accuracy_score, classification_report, confusion_matrix


# Rozdelenie dát na tréningové a testovacie množiny

X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2,
random_state=42)


# Tréning logistickej regresie

model = LogisticRegression()

model.fit(X_train, y_train)
```

 # Predikcia

y_pred = model.predict(X_test)

Výsledky a vyhodnotenie modelu

Po natrénovaní modelu sme ho vyhodnotili pomocou **accuracy, confusion matrix, precision, recall a F1-score**.

 # Vyhodnotenie

accuracy = accuracy_score(y_test, y_pred)

conf_matrix = confusion_matrix(y_test, y_pred)

report = classification_report(y_test, y_pred)

print(f"✅ Presnosť modelu: {accuracy:.2f}")

print("\n📊 Maticová konfúzia:\n", conf_matrix)

print("\n 🔹 Report klasifikácie:\n", report)

Kľúčové zistenia

- **Presnosť modelu: 55 %**, čo naznačuje, že základný model má obmedzenú výkonnosť.
- **Recall pre zdravých pacientov (91 %)** bol vysoký, ale model mal **slabé výsledky pri rozpoznaní vážnejších foriem ochorenia**.
- **Možné vylepšenia** zahŕňajú **použitie iného modelu (Random Forest, SVM, XGBoost), vyváženie datasetu alebo selekciu relevantných premenných**.

Možné vylepšenia modelu

1. Použiť iný model – Logistická regresia je jednoduchý model, ale možno by bol vhodnejší **Random Forest alebo XGBoost**.

2. Zlepšiť distribúciu dát – Niektoré triedy sú nedostatočne zastúpené, preto by sa dalo použiť **Oversampling (SMOTE)** alebo **váhovanie tried**.

3.Redukcia počtu premenných – Použiť **PCA alebo SelectKBest**, aby sa odstránili menej významné atribúty.

Použité technológie

- **Python knižnice (** pandas, numpy, scikit-learn, matplotlib, seaborn)
- **Machine Learning** (logistická regresia, škálovanie, feature selection)
- **Vizualizácie** (scatter ploty, korelačné matice, distribučné grafy)

Záver

Tento projekt demonštruje kompletný postup **dátovej analýzy, predspracovania údajov a modelovania srdcových ochorení pomocou logistickej regresie**. Napriek obmedzeniam základného modelu nám poskytuje cenné poznatky o tom, ako môžeme ďalej zlepšiť predikciu zdravotného stavu pacientov.
